Clinical notes for improved health literacy

Shreyesh Doppalapudi
Department of Information Science, The Pennsylvania State University, Malvern, Pennsylvania, USA
Tingyan Wang
Nuffield Department of Medicine, University of Oxford, Oxford, UK, and
Robin Qiu
Department of Information Science, The Pennsylvania State University, Malvern, Pennsylvania, USA

Abstract

Purpose – Clinical notes typically contain medical jargons and specialized words and phrases that are complicated and technical to most people, which is one of the most challenging obstacles in health information dissemination to consumers by healthcare providers. The authors aim to investigate how to leverage machine learning techniques to transform clinical notes of interest into understandable expressions.

Design/methodology/approach – The authors propose a natural language processing pipeline that is capable of extracting relevant information from long unstructured clinical notes and simplifying lexicons by replacing medical jargons and technical terms. Particularly, the authors develop an unsupervised keywords matching method to extract relevant information from clinical notes. To automatically evaluate completeness of the extracted information, the authors perform a multi-label classification task on the relevant texts. To simplify lexicons in the relevant text, the authors identify complex words using a sequence labeler and leverage transformer models to generate candidate words for substitution. The authors validate the proposed pipeline using 58,167 discharge summaries from critical care services.

Findings – The results show that the proposed pipeline can identify relevant information with high completeness and simplify complex expressions in clinical notes so that the converted notes have a high level of readability but a low degree of meaning change.

Social implications – The proposed pipeline can help healthcare consumers well understand their medical information and therefore strengthen communications between healthcare providers and consumers for better care.

Originality/value – An innovative pipeline approach is developed to address the health literacy problem confronted by healthcare providers and consumers in the ongoing digital transformation process in the healthcare industry.

Keywords Health literacy, Information extraction, Multi-label classification, Lexical simplification, Word embedding, BioBERT, LSTM, Transformer models, Pre-trained language models, Natural language processing, NLP

Paper type Research paper
Introduction

In accordance with US government’s Healthy People 2030 initiative (NIH, 2020), personal health literacy is about an individual’s ability to find, understand and use information for health-related decisions and actions, while organizational health literacy concerns the degree to which organizations enable individuals to enforce personal health literacy. Both personal and organizational health literacy are essential for information exchange between healthcare consumers and providers, which is crucial for proper care and use of services and for patients to make decisions and take actions. Low health literacy can negatively affect patient care and outcomes and healthcare utilization (Berkman, Sheridan, Donahue, Halpern, & Crotty, 2011). Limited health literacy happens when individuals’ literacy and numeracy skills are mismatched with the information that organizations make available. The Program for the International Assessment of Adult Competencies (PIAAC) reported that only 14% of the US adult population was scored in the highest literacy proficiency level, 10% in the highest numeracy proficiency level and 6% in the highest digital skill proficiency level (PIAAC, 2017).

Each of these skills are important components for developing health literacy, as these skills are required to find, understand and use health information and services. A lower measure in any of these skills directly correlates to lower health literacy rates.

The benefits of higher health literacy include more effective communications, better adherence to treatment, greater ability to engage in self-care, and therefore leading to improved healthcare outcomes and reduced healthcare cost (Chang, 2011; Morrison, Glick, & Yin, 2019). To improve health literacy, it requires healthcare providers to avoid complex and jargon-filled language in disseminating health information (Hersh, Salzman, & Snyderman, 2015) in addition to improving consumers’ literacy, numeracy and digital problem solving skills. The solutions for both these requirements are long-drawn and complicated for both the care providers and the consumers. Promisingly, automated solutions using natural language processing (NLP) techniques and machine learning (ML) methods can help bridge the gap between both sides and hence provide more opportunities for better care (Hendawi, Aliaan, & Li, 2022). It is well known that clinical notes represent a huge collection of information on patients, including the whole process of caregiving ranging from patients’ diagnosis and admission to discharge. To promote health literacy, consumers need to derive the maximum value out of clinical notes, requiring the ability or tools to process health-related information in their medical notes. For these unstructured digital clinical notes, NLP- and ML-based methods can be used to identify the information of interest and simplify specialized expressions to help patients with a better understanding of their clinical notes. On the other hand, this frees up time and effort for the care providers which, in turn, can be spent toward care of patients instead of administrative tasks.

For the last decades, the rise of deep learning (DL), a specialized subset of ML, has provided a new lease toward the field of NLP (Miotto, Wang, Wang, Jiang, & Dudley, 2018). Moreover, pretrained language models over last few years have driven the NLP field into a new era (Wang, Xie, Pei, Tiwari, & Li, 2021). Higher storage and computing power allow for large-scale models, leading to better results on various NLP tasks. Information extraction and text simplification are two of the most important NLP tasks that support health care consumers to understand and harness the complete information from their own medical notes. These two NLP tasks therefore have the piqued interest of a variety of research communities to develop optimal outcomes for the consumers.

The opportunity for automated solutions based on NLP and ML methods to bridge the gap in information dissemination from healthcare providers to consumers drives this study (Doppalapudi, 2021). The study is to focus on extracting relevant information and simplifying medical jargons from long and unstructured digital clinical notes. We therefore set out to answer the following key questions: (1) which NLP mechanism is required to identify and extract relevant information from long clinical narratives? (2) Which ML based process can be used to verify the completeness of the extracted information? (3) Which NLP technique is
required to simplify text by identifying and replacing medical jargons? And (4) what metrics can be used to automatically evaluate the readability of the simplified notes while preserving the original meaning.

**Related work**

Researchers have solved the task of extracting relevant text from medical text using a variety of different methods. Various ML and NLP techniques such as Naïve Bayes classifier, support vector machine, convolutional neural network (CNN) (Tran & Kavuluru, 2017), recurrent neural network (Liu, Tang, Wang, & Chen, 2017), attention model (Gao et al., 2018), topic modeling (Rumshisky et al., 2016), rule based model (Weissman et al., 2016; Wulff et al., 2020), hybrid model with a combination of rule-based and ML methods (Byrd, Steinhubl, Sun, Ebadollahi, & Stewart, 2014; Chen, Song, Shao, Li, & Ding, 2019) and transfer learning (Giorgi & Bader, 2018) were used heavily in previous approaches. Of all the ML techniques, DL models have provided better results.

Classification has been the most popular method to verify the effectiveness of the text extractions, providing with easier evaluation options. Previous research studies have also attempted to develop methods for classification of clinical notes directly into international classification of diseases (ICD) codes using Medical Information Mart for Intensive Care (MIMIC) database (Johnson et al., 2016), which is a popular database used to predict the ICD codes associated with the clinical notes archived in the database. Particularly, it provides an opportunity for multi-label classification modeling on the real-world data from critical care. Different approaches have been proposed for disease code classification task, such as DL models (Li et al., 2019; Hsu, Chang, & Chang, 2020) and topic modeling (Gangavarapu, Jayasimha, Krishnan, & Kamath, 2020). In addition, clinical notes in other languages can be classified in similar models which has been popular specifically in Spanish (Perez, Perez, Casillas, & Gojenola, 2018; Blanco, Perez-de-Vinaspre, Perez, & Casillas, 2020; Almagro, Unanue, Fresno, & Montalvo, 2020), and the approaches proposed have been similar to that used for clinical notes in English.

Earlier version of lexical simplification involved the identification of complex entities to provide remedial measures on complicated concepts and simplification of those concepts. These complicated concepts include negated concepts, abbreviations and composite and implicit entities. A series of studies had focused on dealing with each of these complex entities separately and proposed varying remedial information retrieval methods to provide the relevant information and simplified concepts, including negation type classification (Mukherjee et al., 2017), negated concept detection (Peng et al., 2018), abbreviation disambiguation (Joopudi, Dandala, & Devarakonda, 2018), implicit entity recognition (Perera et al., 2015) and composite entity components identification (Wei, Leaman, & Lu, 2015). In addition to entity recognition, studies on nonmedical text lexical simplification provide different approaches for the task, which can be adapted to the task on biomedical text. Neural text simplification models’ variants are common in research with non-medical text, along with the evaluation metrics used for quantifying the performance of these models (Demirtas, Cicekli, & Cicekli, 2010; Cer, Manning, & Jurafsky, 2010; Nisioi, Stajner, Ponzetto, & Dinu, 2017; Qiang, Li, Zhu, Yuan, & Wu, 2020).

Generally, the lexical simplification task can be split into two steps, i.e. complex word identification and substitute candidate generation. For the former, previous studies proposed varying methods ranging from rule-based methods to word embeddings (Maddela & Xu, 2018; Pylieva, Chernodub, Grabar, & Hamon, 2018; Alfano et al., 2020). For the substitute candidate generation, researchers have used phrase tables to link complex medical terms to simple laymen phrases or words (Chen et al., 2018; Shardlow & Nawaz, 2019). Research into the field of medical text simplification has gathered steam over the last few years. The rise of DL through the increase in available computational power, development of hierarchical attention models, and advances in large scale NLP systems have allowed for more research in
medical text simplification over the past few years. Research focus therefore has been shifted from rule-based models to attention models for the text simplification task (Moradi & Ghadiri, 2018; Van den Bercken, Sips, & Lofi, 2019; Kauchak & Leroy, 2020; Sakakini et al., 2020; Van, Kauchak, & Leroy, 2020; Li et al., 2022).

As shown from the known research approaches, each of the fields related to text extraction, text classification and lexical simplification using clinical notes is a relatively new topic of interest. Regarding text extraction, existing research has focused on modeling the problem as name entity recognition (NER) task of extracting just words or phrases most relevant to entity. Our study augments the existing approach by adding a text summarization directive to identify most relevant sentences to the entity and then filter them to generate a summary. We aim to improve the performance of these text extraction models by leveraging the power of word embeddings to create a target vocabulary for an entity (in this case, a disease) and filtering based on keyword mapping to this vocabulary.

No automated methods exist for directly answering the question of the relevant text information, especially in the absence of parallel corpora to train the model on. Therefore, we select the multi-label classification of the extracted notes into ICD codes as the evaluation process for relevant text extractor. The classification is performed for the most common 50 and 100 labels with standardized labels for 3-digit and 4-digit codes. Note that classification into ICD codes has never been attempted to prove the validity of data itself. Simply, we aim to achieve parity in performance with the state-of-art to show that no vital information with respect to diagnosis will be lost while extracting the relevant information.

For the lexical simplification research, we aim to leverage the power of the transformer models pre-trained on large scale medical text data along with embedding, which has not been attempted in the field yet. Although the approach of text simplification with transformers has been attempted on non-medical text, its performance is still lagging on medical notes since medical notes have a whole new dimension of difficulty and complex terms along with jargons and abbreviations that are very specific to the medical domain. These complexities are easier to qualify by medical experts, however, to evaluate quantitatively requires different approaches. We consider combining the evaluation approach of using readability indices on a document-level simplification and machine translation metrics to understand the change of grammar and meaning in sentences during lexical simplification. This combination of metrics can provide a robust automatic evaluation process to allow for better speed in model development. This evaluation process will provide means to iterate on model versions faster and allow human interpretation at the final stage to understand the overall performance of the model.

Methodology

We formulate the information extraction and simplification for clinical notes as a pipeline problem with three-stage tasks: Stage I is to extract relevant information of interest, Stage II is to map the relevant information into patients’ diagnosis codes for information completeness check, and Stage III is to simplify the information extracted. Therefore, we proposed an NLP pipeline that can extract relevant information and simplifying lexicons from long unstructured clinical notes, as shown in Figure 1. For extracting a relevant text (Stage I), we develop an unsupervised keyword matching method to extract diagnosis information from clinical notes, wherein similar word vocabulary for each target diagnosis is created using a pre-trained word embedding. To automatically evaluate the completeness of the extracted information, we perform a multi-label classification task on the relevant text (Stage II), with a comparison to the state-of-art results. For text simplification, we identify and mask complex words in the text extracted and then generate candidate words for the masked positions (Stage III). In the following text, we will introduce the details of methods development for each stage.
Stage I: Relevant text extraction

To develop an information extraction model that can capture all relevant mentions of disease information contained in clinical notes, we would need an entity filter that is trained on a large corpus of medical text to identify the most used words and phrases directly correlated to a disease as well as other similar words and abbreviations associated with the disease. For this purpose, we consider using a word embedding trained on a large corpus in medical domain to find similar words of disease descriptions and then build a target vocabulary for each type of disease. In this study, we use the pre-trained word embeddings provided by (Pyysalo, Ginter, Moen, Salakoski, & Ananiadou, 2013), which were trained on abstracts and all full-text documents from PubMed Central Open Access subset, a database hosted by the National Institute of Health (NIH).

We tokenize long form descriptions of ICD-9 codes into words for each disease, subsequently moving the stopwords and generating a keyword set. Then using the pre-trained embeddings we identify similar words to the keywords for each disease based on cosine distance with a threshold value. As a result, a target vocabulary is generated for each disease, which consists of the similar words and the keywords (Figure 2a). Typically, if a sentence mentions words and phrases from the target vocabulary, it indicates the presence of relevant information in the sentence for that particular disease.

To extract relevant sentences from long clinical notes, we design an unsupervised regex-based keywords matching filter. The filter retains information about medical history and diagnoses related to the patient but exclude the social history information or other sections that do not contain any medical information. We tokenize clinical notes into sentences and then use the matching filter to select relevant sentences, which are then joined together to form the relevant text of the clinical note. The process of extracting target/relevant sentences is shown in Figure 2b.

Stage II: Text classification (mapping relevant information to diagnosis codes)

To evaluate the completeness of the extracted information, we map relevant information to patients’ diagnosis codes recorded for the same hospital stay. We develop a multi-label classification model based on long short-term memory (LSTM) units to perform this task (Figure 3a) and compare the model performance to the state-of-art. Specifically, we use ICD codes as true labels that were assigned at the same hospital stay corresponding to the clinical notes extracted. Figure 3b shows the broad outline of the model building process in Stage-II. The preprocessing of notes involves the removal of stopwords and lemmatization to obtain root words. Then the vocabulary size is determined to define the size of the embedding layer using Keras Tokenizer API, which also helps in determining the sequence length distribution for set of notes. Finally, vectorization is performed on the pre-processed notes to convert them into sequences and the sequences are padded to the same length. In this study, we use top 50
or 100 ICD codes for modeling, which are selected on the basis of the codes with higher number of admissions. Both 3-digit and 4-digit ICD codes are considered when selecting the top-50 or -100 ICD codes, but a different set for each type of codes is considered exclusively. Specifically, only 3-digit codes are used in the set of 3-digit codes and 4-digit codes in the set of 4-digit codes.

At this stage, F-1 score and accuracy are utilized as the evaluation metrics. For utilizing these metrics effectively for the multi-label classification problem, we use micro-averaging method. Micro-averaging method allows emphasis on the performance on more common labels and does not allow high performance on rare labels to affect the final evaluation score. For comparison, we utilize the results from Hsu et al. study (Hsu et al., 2020) as the state-of-art, particularly the results from their best-performing CNN models. We compare the performance of these existing models with our proposed classification models at Stage II across top-50 and top-100, 3-digit and 4-digit ICD codes using the micro-averaged F-1 score as the evaluation metrics.

**Stage III: Lexical simplification**

We perform lexical simplification by replacing complex words in the relevant text extracted with simpler alternatives of equivalent meaning. Lexical simplification includes a few sub-tasks that need to be performed in sequence, including complex word identification (CWI), similar word generation, ranking filtering and substitution. First, we use a sequence labeler
model (SEQ) (Gooding & Kochmar, 2019) to identify complex words in text and then masked these identified words (Figure 4). SEQ is built on bi-directional LSTM units to allow for context to be learned around the target word, which considers both context of word and morphological structure of word while identifying the complex word.

We then leverage BioBERT based transformer models to generate candidate words for the masked positions, subsequently ranking and filtering the candidate words for substitution (Figure 4). BioBERT model is pre-trained on large scale biomedical corpora (Lee et al., 2020), which outperforms BERT on three representative biomedical text mining tasks. BioBERT models can be used to predict the word of the masked position by considering the context around mask. We utilize two different versions of BioBERT, i.e. base version (v1.0) and large version (v.1.1). BioBERT-large was trained on one million PubMed abstracts with a vocabulary size of around 30,000 while BioBERT-base was trained on 200,000 PubMed abstracts and 270,000 PubMed Central full length texts with a vocabulary size of around 29,000 (Lee et al., 2020). As BioBERT was originally trained for NER task, it requires repurpose to perform lexical simplification. Specifically, we repurpose the intermediate fully
connected layer from the encoder to fine-tune the representation of lexical simplification. In
addition, we use different embeddings to achieve the best possible simplification results. One
is the embedding directly available in the BioBERT model with a dimension of 512, while
another is obtained from the study by Pyysalo et al. with a dimension of 400 (Pyysalo et al.,
2013). Pyysalo et al.'s embeddings are trained on PubMed abstracts using Word2Vec training
process allowing more contextual information to be represented in the embedding while
BioBERT’s embedding is based on the frequency of occurrence of the word in the corpus.
To rank the candidate words predicted, we use both the Zipf frequency (Li, 1992) and
Bilingual Evaluation Understudy Score (BLEU) (Papineni, Roukos, Ward, & Zhu, 2002), a
machine translation metric. Zipf frequency is the base-10 logarithm of the number of times it
appears per billion words. If the zipf frequency is high, the word is more commonly used in
texts, which means it is easier for most people to understand. BLEU is a metric to make sure
that the output sentence is similar in meaning and grammar to the original sentence (before
masking complex words), which is used as a reference similar to machine translation
situations. Masked Language Modeling (MLM) Likelihood (Devlin, Chang, Lee, & Toutanova,
2018) is used as the loss function and Adam is used as the optimizer with a 0.0001 learning rate.
At this stage, readability and the degree of change are used to evaluate the lexical
simplification task. We use three different readability indices including Flesch–Kincaid
(Kincaid, Fishburne, Rogers, & Chissom, 1975), Gunning–Fog (Gunning, 1952), and Coleman–
Liau (Coleman & Liau, 1975), the mathematic calculations of which are provided in Equations
(1) to (3) below:

\[
FK = 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right) 
\]

\[
GF = 0.4 \left[ \left( \frac{\text{number of words}}{\text{number of sentences}} \right) + 100 \left( \frac{\text{number of complex words}}{\text{number of words}} \right) \right] 
\]

\[
CLI = 0.0588(L) - 0.296(S) - 15.8 
\]

To evaluate the degree of change, we use BLEU (Papineni et al., 2002) to test the changes to
text using words from reference sentences, with additions and deletions penalized.

Data sources and preprocessing
In this study, we use data from Medical Information Mart for Intensive Care III (MIMIC-III)
database version 1.4, in which data were collected from critical care services, comprising de-
identified health data associated with ∼58,000 intensive care unit admissions (Johnson et al.,
2016). MIMIC-III contains information around diagnosis, laboratory tests, medications,
procedures, vital signs and caregiver information for each patient admission. For validating
our proposed NLP pipeline, we consider using discharge summary notes as they contain the
most comprehensive information around patients’ hospital stays, including illness history,
diagnosis, medication and test results. Totally, 58,167 discharge summary notes are retrieved
from NOTEEVENTS table in the database for this study.
To facilitate verifying the completeness of relevant information extraction, we also retrieve all the diagnosis codes from DIAGNOSES_ICD table assigned for the same hospital stay corresponding to the discharge summary notes for each patient. We select top 50 and 100 ICD codes with higher number of admissions in the MIMIC III database, in terms of 3-digit and 4-digit ICD codes respectively. We also use the definitions of ICD Version 9 (ICD-9) codes for diagnoses from D_ICD_DIAGNOSES table to build our vocabulary. ICD-9 codes have a hierarchical representation with three-digit codes being top rung and 4th and 5th digits forming the sub-section of diseases. The table consists of 14,567 different ICD-9 codes along with a short and long description for each code. Typically, these codes are assigned at the end of a patient’s stay and are used by the hospital to bill for care provided. Therefore, these ICD-9 codes along with descriptions are used for creating target vocabulary toward relevant information extraction.

We join NOTEVENTS table and DIAGNOSES_ICD table using hospital admission id (HADM_ID). Basic checks on the data such as identification and removal of missing values, ensuring consistency in time periods of notes generation with time periods for other parts of the database are performed before utilizing the data for modeling purposes. Particularly, quality checks on data are conducted by examining any empty notes and verifying the chart time is before the store time. The chart time for a clinical note (the time when the note is created) must be before the store time (the time when the note is stored) for any valid entry of note. These quality checks reveal no other note and all the discharge summary notes mentioned previously are retained. After the data cleaning process has been completed the text part of the discharge summaries are used as the input to the relevant text extraction model at Stage I.

Results and findings
Sets of similar words are generated using the pre-trained embeddings, results of a few representative word clusters for myocardial infarction, diabetes, cancer or hypothyroidism are provided in Supplementary Figure A1. The results show that a similar subset covers a lot of other downstream tasks in the entity filter such as implicit entity, negation, complex entities while also considering a few spelling mistakes. Based on the target vocabulary, relevant sentences are extracted by the built entity filter. Figure 5 illustrates results of the relevant information extraction process on two discharge summary notes. The filter retains information on the medical history and the diagnosis provided to the patient, while excludes the social history information that is not directly relevant to any of the disease diagnosis code, or sections that do not contain any medical information.

The extracted notes obtained from Stage I are further processed with the removal of stopwords and lemmatization to obtain root words for determining vocabulary size to define the size of embedding layer in the classification model. With vectorization is performed on the processed notes to convert them into sequences and the sequences are padded to the length of 2,000. The sequence length of 2,000 is selected because it covered a majority of the notes in terms of notes length as shown in the distribution of sequence length of notes vs. the number of notes in Figure 6. These padded sequences are passed on as input to an untrained embedding layer in the classification model. The architectures employed to train the DL model for the multi-label classification tasks on Top-50 and Top-100 ICD codes are provided in Supplementary Tables A1 and A2, respectively.

Table 1 provides the performance evaluation for the classification task along with the comparison with the state-of-art research using the micro-averaged F-1 score metric, while Table 2 provides the micro-accuracy of different models. As shown in both tables, the performance of the proposed classification models is close with the state-of-art CNN models, though there still some differential in performance to be achieved from Stage-II models. This difference in the performance level can attributed to the disparity of the input notes for the
### Figure 5.

Results of the information extraction in Stage-I. All medical and diagnosis information is retained while social history (the example in panel a) or sections with no meaningful information are dropped (the example in panel b).

### Figure 6.

Distribution of sequence length vs. number of notes.
models while training. This might indicate that there is still some information that is dropped during Stage-I process, nevertheless Stage-I is still robust to a large extent as evidenced by the small differential in performance to the state-of-art.

In comparison, the performance differential is lower for the top-100 ICD codes models in comparison to the top-50 ICD code models. There is a difference in performance between 3-digit and 4-digit models. This is expected as the 4-digit codes are more specific in comparison with 3-digit codes. For example, ICD-9 code 428 represents Heart Failure while 428.2 and 428.3 represent Systolic and Diastolic Heart Failure, respectively. Hence, it is easier to classify accurately from clinical notes for more specific ICD codes especially when the label ICD codes are standardized. In contrast, this is flipped in the F-1 score performance because of more breadth of information allowed inside 3-digit codes that allow for more reasonable misclassifications and result in a better performance while combining precision and recall.

Tables 3 and 4 present the performance of the lexical simplification model among the various experiments with different combinations of transformer models, ranking mechanisms and word embeddings, evaluated by readability indices and degree of change metrics, respectively. In terms of readability indices, we compare the results to the state of the art results from the study by (Shardlow & Nawaz, 2019) that worked with biomedical text. For the degree of change metrics, we compare them to the results from the study by (Nisioi et al., 2017) that worked on general domain text. As shown in both tables, the BioBERT-Large model provides better performance in comparison to the BioBERT-Base model. The experiments with the BioBERT-Large have consistently better scores on all the readability indices and both BLEU and SARI metrics. There is an average difference of 2 on the Gunning–Fog reading grade, 20 on BLEU score and 4 on SARI between the experiments on BioBERT-Base and BioBERT-Large models. This can be attributed to the fact of a larger pre-training corpus for the BioBERT-Large model. BioBERT-Large model works with a larger
vocabulary and the weights are trained on a larger corpus of biomedical text representing more relationships among words.

Text simplification results for three example paragraphs, showcasing the output from the Stage-III lexical simplification model, are provided in Table 5, where complex words and their

<table>
<thead>
<tr>
<th>Degree of change metrics</th>
<th>BLEU</th>
<th>SARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioBERT Base-Z</td>
<td>28.72</td>
<td>14.12</td>
</tr>
<tr>
<td>BioBERT Base-Z + BLEU</td>
<td>51.36</td>
<td>18.29</td>
</tr>
<tr>
<td>BioBERT Base-Z + BLEU + Pyysalo</td>
<td>67.61</td>
<td>21.57</td>
</tr>
<tr>
<td>BioBERT Large-Z</td>
<td>56.49</td>
<td>18.97</td>
</tr>
<tr>
<td>BioBERT Large-Z + BLEU</td>
<td>75.89</td>
<td>25.93</td>
</tr>
<tr>
<td>BioBERT Large-Z + BLEU + Pyysalo</td>
<td>79.96</td>
<td>27.68</td>
</tr>
<tr>
<td>State-of-art performance (Nisioi et al., 2017)</td>
<td>87.50</td>
<td>31.11</td>
</tr>
</tbody>
</table>

Table 4. Results of stage III lexical simplification (metrics for degree of change)

Original text                                      | Simplified text |
---------------------------------------------------|-----------------|
Brief Hospital Course: Patient presented electively for meningoia resection of [**3-5***]. She tolerated the procedure well and was extubated in the operating room. She was transported to the ICU post-operatively for management. She had no complications and was transferred to the floor and observed for 24 hours. Prelim path is consistent with meningoia. She has dissolvable sutures, and will need to come to neurosurgery clinic in [**6-28***] days for wound check only. She will need to be scheduled for brain tumor clinic. She will complete Decadron taper on [**3-10***] and then restart her maintenance dose of prednisone. She will also be taking Keppra for seizure prophylaxis. Her neurologic examination was intact with no deficits at discharge. She was tolerating regular diet. She should continue to take over the counter laxatives as needed.

CXR [**2125-2-9***]: The patient is after median sternotomy and CABG. Bilateral perihilar haziness continues toward the lower lungs is new consistent with new moderate-to-severe pulmonary edema. Bilateral pleural effusion is present, also new, most likely part of the heart failure. Left and right retrocardiac opacities consistent with atelectasis.

3. Coronary artery disease: Patient with a history of myocardial infarction in [**2180***] and [**2182***] and is status post stent of the percutaneous transluminal coronary angioplasty in [**2182***]. Enzymes were cycled, which were negative. Aspirin and Coumadin were held due to gastrointestinal bleed. Beta blocker and ace were initially held due to low blood pressures. Lipitor was held secondary to new cirrhosis. The patient was restarted on Nadolol upon discharge, however, aspirin, Coumadin, Zestril and Lipitor were held prior to discharge to be restarted by primary care physician at his or her discretion.

Table 5. Results of simplifications generated by the lexical simplification model
substitutions are highlighted in grey shading. As can be seen from these examples, the lexical simplification models simplify complex terms while ignoring non-replaceable terms such as surgery names, medicines and marginal disease names. The substitutions maintain the meaning of the original sentence in most cases, while providing more readable text for an average user.

Discussions
Improving health literacy is a responsibility for both the organizations and the consumers. The biggest roadblock in the path to improved health literacy is the ability of consumers to read and assimilate information from clinical documents. We proposed a novel framework based on NLP and DL techniques to help consumers with a better understanding of the information in their digital clinical notes. The three important stages (i.e. information extraction, verification and simplification) in the proposed framework involve different but complementary types of tasks. This framework provides a great benefit to both healthcare consumers and providers, as it allows for improving health literacy by working on different facets of the problem related specifically to the digital clinical notes. The combination of relevant text extraction and lexical simplification allows consumers to understand and process their health information from their electronic health records, specifically clinical notes by limiting the burden of jargons and writing style while preserving the original meaning. Besides acting as performance evaluation role for Stage I, the multi-label classification model in Stage II also allows organizations to enforce the health literacy policies by verifying the information contained in the records automatically, without putting more burden on the caregivers potentially avoiding them legal and compliance issues while allowing them to fulfill their duty in promoting health literacy among their consumers.

The methodological contributions of this study can be summarized as follows: (1) a method for extracting relevant text from long clinical narratives, with building a comprehensive target vocabulary based on the descriptions of diagnosis codes and word embeddings trained on large corpus, (2) a multi-label classification model trained on Top-50 and Top-100 ICD codes, allowing for an automatic verification on information completeness of the upstream task (relevant text extraction) via text classification, (3) a transformer-based lexical simplification model utilizing contextual word embeddings and machine translation metrics in ranking mechanism, and its performance is comparable to the performance of models developed in previous studies for a smaller data set or general text with a lower degree of complexity and (4) an approach and metrics to automatically evaluate the performance of lexical simplification model from different perspectives including readability and degree of change. Overall, the novel design of stages in the proposed pipeline allows for transforming long unstructured clinical notes for improved health literacy via text extraction with minimal information loss and text simplification with low degree of change. In addition, the performance of the classification models at Stage II indicates that the extraction model at Stage-I retrieve high quality of relevant information with an uncomplicated methodology along with being space- and time-efficient algorithm. This can help with providing real-time support to the consumers when required, as the algorithm is fast with reasonable results. This is one of the drawbacks of existing research approaches which are space- and time-exhaustive in search for high accuracy leading to resource scarcity and higher costs of implementation.

As for the text simplification at Stage III, the addition of BLEU score to the ranking measure has boosted the performance of the models. The boost is consistent across all readability indices and degree of change metrics. This reveals that models incorporated with BLEU help not only make sure that the output sentence and original text are as close as possible, but also is able to filter better candidates for achieving higher readability indices scores. It is observed...
that the effect of BLEU as a ranking measure is lower on the readability indices in comparison with the degree of change metrics. In addition, Pyysalo et al.'s (2013) embedding boosts the model performance and turns out to be the best-performing models across both BioBERT-Base experiments and BioBERT-Large experiments. The context-aware embedding has improved similar word generation, which in turn results in better candidate substitutions. Especially, the models with the ranking mechanism of a combination of Zipf frequency and BLEU score outperform the embedding with the original BioBERT models.

There are however some limitations that should be pointed out. The cosine similarity threshold is a difficult parameter to estimate. To test it for different values of the parameter, the whole process of Stage-I and Stage-II need to be performed, which is time-intensive. To test the effect of different cosine similarity thresholds on the relevant text extraction process will be a difficult proposition. Also, the performance of each stage in this study is dependent heavily on pre-trained word embeddings selected. Large embeddings allow for more word-to-word relationship representation, but it is difficult to know the tradeoff between the computational complexity of training using embeddings vs. the benefit of performance in advance which again leads to a lot of time intensive and computational power intensive work. Models at Stage-II in this research have performances close to the state-of-art, yet not surpassed it. The reason might be that the input to the Stage II classification model (the text extracted at Stage I) proved to drop a little relevant information. Nevertheless, this still shows that the Stage-I relevant text extraction is robust enough to provide a good performance. The best-performing models at Stage III comparable to but did not outperform the state-of-art performance, in terms of both the readability indices and the degree of change metrics. From the perspective of readability indices, Shardlow and Nawaz dealt with a subset of only 500 discharge summary notes (Shardlow & Nawaz, 2019) whereas our study has worked on a far larger data set – around 58,167 discharge summary notes. Depending on the complexity of the documents in the sample, the readability indices score might change significantly. Regarding the degree of change metrics, it is generally explained that as the state-of-art was trained on general text rather than biomedical text. Typically, the complexity of general text is quite low and has a lower percentage of complex words compared to that of biomedical text where jargons and specialized words is much more commonly seen (Rothrock et al., 2019). It therefore is easier to perform lexical simplification on general text with a lower degree of change as explained by BLEU score and provide effective outputs as characterized by the SARI score. Although the transformed clinical notes might lose some information compared to the original text using medical jargon and specialized words, the proposed transformation pipeline helps improve health literacy substantially, resulting in information gain for individual patients when compared with the acquired information by reading and comprehending the original clinical records. Promisingly, the proposed pipeline can be further improved by incorporating knowledge graphs (Hendawi et al., 2022) and biomedical vocabularies and ontologies repository, e.g. Unified Medical Language System (Bodenreider, 2004) for target vocabulary enrichment and candidate substitutions generation, and also by training the simplification model with larger data sets in medical domain.

In our future research, the creation of target vocabulary can be further investigated, for example, extending to phrases with the use of n-grams for similar word extraction and relevant sentence filtering. Similar to Stage-I, in Stage-III, we have only considered word-for-word replacement in terms of text simplification. Note that phrases simplification is required for lexical simplification along with the word-for-phrase and phrase-for-word replacements to allow for best performing simplification models. In addition, lexical simplification is just a sub-task of text simplification, where content reduction is another sub-task of text simplification. Transformer models can be repurposed to perform both lexical simplification and content reduction for text summarization, but evaluation cannot be exclusively automatic and need human expert evaluation in that case. The application in the real world of
Stage II with the multi-label classification model can be further explored, e.g. to evaluate the comprehensiveness of notes produced by healthcare providers. With enough confidence in the performance of the classification model, it can be utilized to cross-check diagnosis notes with the ICD codes in billing, providing avenues to check for compliance and avoiding potential legal issues in the future. Moreover, it frees up the time for the healthcare providers to involve themselves in more clinical work instead of being tangled in an administrative process. Last but not least, besides using readability to evaluate clinical notes transformed it would be useful to develop a set of measures of effectiveness to further prove the improvement of health literacy with respect to these measures, which can be designed based on Patient and Public Involvement and Engagement programs. Promisingly, the methods within the proposed framework in this study can be extended to clinical documents in other languages if there are pre-trained word embeddings and pre-trained transformer models trained on medical text in that language.

Conclusions
This study proposed a multi-stage NLP pipeline of relevant text extraction, verification and simplification in digital clinical notes. In the proposed NLP pipeline, the keyword-based entity map filter for relevant information extraction was built based on a similar word construction from a word embedding. For verifying the completeness of the information extracted, a multi-label classification task on the most common labels was performed which allows for a comparison with the literature. This is based on the hypothesis that if the model performs close to the state-of-art, then most of the relevant information is retained. Finally, a lexical simplification method was developed, which consists of a sequence labeler and transformer-based models, with the former for identifying and masking complex words in the text extracted and the latter for substituting the words masked. The performance of lexical simplification method was evaluated from two perspectives – how simple the text has been converted into and the amount of meaning and grammar lost to achieve that level of readability. More importantly, we utilized discharge summary notes from MIMIC-III data set to validate our proposed framework.

With the fast development of information technologies, electronic medical records have been widely used by most healthcare providers. This is particularly true in the developed world. Today, patients can view their medical record through their service providers’ portals. If this multi-stage NLP pipeline of relevant text extraction, verification and simplification is adopted in practice, it will potentially help patients better understand their health information in clinical notes with as little help from their providers as possible. Therefore, the developed approach will contribute to addressing the health literacy problem confronted by healthcare providers and consumers in the ongoing digital transformation process in the healthcare industry.

References


Supplementary materials

Figure A1. Results of similar words for (A) cancer, (B) diabetes, (C) myocardial infarction or (D) hypothyroidism. The word is highlighted in the blue rectangle while its closest words are provided in the list.
Table A1. An example of the classification model architecture (top-50 ICD codes)

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output shape</th>
<th>Param #</th>
</tr>
</thead>
<tbody>
<tr>
<td>embedding (Embedding)</td>
<td>(None, 2000, 128)</td>
<td>30,189,696</td>
</tr>
<tr>
<td>lstm (LSTM)</td>
<td>(None, 2000, 256)</td>
<td>394,240</td>
</tr>
<tr>
<td>dropout (Dropout)</td>
<td>(None, 2000, 256)</td>
<td>0</td>
</tr>
<tr>
<td>lstm_1 (LSTM)</td>
<td>(None, 2000, 128)</td>
<td>197,120</td>
</tr>
<tr>
<td>dropout_1 (Dropout)</td>
<td>(None, 2000, 128)</td>
<td>0</td>
</tr>
<tr>
<td>lstm_2 (LSTM)</td>
<td>(None, 64)</td>
<td>49,408</td>
</tr>
<tr>
<td>dropout_2 (Dropout)</td>
<td>(None, 64)</td>
<td>0</td>
</tr>
<tr>
<td>dense (Dense)</td>
<td>(None, 50)</td>
<td>3,250</td>
</tr>
</tbody>
</table>

Note(s): Total params: 30,833,714; Trainable params: 30,833,714; Non-trainable params: 0

Table A2. An example of the classification model architecture (top-100 ICD codes)

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output shape</th>
<th>Param #</th>
</tr>
</thead>
<tbody>
<tr>
<td>embedding (Embedding)</td>
<td>(None, 2000, 128)</td>
<td>30,574,464</td>
</tr>
<tr>
<td>lstm (LSTM)</td>
<td>(None, 2000, 256)</td>
<td>394,240</td>
</tr>
<tr>
<td>dropout (Dropout)</td>
<td>(None, 2000, 256)</td>
<td>0</td>
</tr>
<tr>
<td>lstm_1 (LSTM)</td>
<td>(None, 2000, 128)</td>
<td>197,120</td>
</tr>
<tr>
<td>dropout_1 (Dropout)</td>
<td>(None, 2000, 128)</td>
<td>0</td>
</tr>
<tr>
<td>lstm_2 (LSTM)</td>
<td>(None, 64)</td>
<td>49,408</td>
</tr>
<tr>
<td>dropout_2 (Dropout)</td>
<td>(None, 64)</td>
<td>0</td>
</tr>
<tr>
<td>dense (Dense)</td>
<td>(None, 100)</td>
<td>6,500</td>
</tr>
</tbody>
</table>

Note(s): Total params: 31,221,732; Trainable params: 31,221,732; Non-trainable params: 0

Corresponding author
Robin Qiu can be contacted at: robinqiu@psu.edu

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com